alineos
High Performance Linux Solutions

cluster de calcul parallèle linux

manuel d'utilisation

ALINEOS SA - Société Anonyme au capital de 45 734,70 € - RCS Montereau B 430 100 628 - APE 725 Z
14 bis, rue du Maréchal Foch - 77780 BOURRON MARLOTTE
Téléphone : 01 64 78 57 65 - Télécopie : 01 64 78 57 66 - E-mail : info@alineos.com - Web : www.alineos.com

INDEX

1 – CONFIGURATION LOGICIELLE DU CLUSTER

2 – PROCEDURE DE DEMARRAGE ET ARRET DU CLUSTER

3 – EXPLOITATION DU CLUSTER

4 – EXPLOITATION LOGICIELLE

5 – MONITEOS v2.6

6 – CHECK LIST D'INSTALLATION CLUSTER

1.1 – CONFIGURATION LOGICIELLE DU CLUSTER : SERVEUR(S)

Distribution Linux	**Red Hat 7.3**
Kernel	**2.4.18 ou sup**

image de la distribution	/opt/RH-7.3

nom de(s) machine(s)	master**$n**.domaine.com n= 0, 1… par défaut : master0.clustal.com

eth0	192.168.1.15**$n**
netmask	255.255.255.0
réseau	192.168.1.0
diffusion	192.168.1.255

eth1	
netmask	
réseau	
diffusion	

partitions des disques durs	
/boot	100 MO
/	10 GO
/home	->
swap	RAM x 2

compte root	mot de passe : xxxxxx
compte admin	mot de passe : xxxxxx

MPI-CH-1.2.4	/usr/local/mpich-1.2.4

LamMPI-6.5.7	/usr/local/lam-mpi

OpenPBS-2.3.16	/usr/local/OpenPBS_2_3_16

ATLAS-3.4.1	/usr/local/ATLAS

LAPACK-3.0	/usr/local/LAPACK

MPIBLACS-1.1	/usr/local/BLACS

PVMBLACS-1.1	/usr/local/BLACS

PVM-3.4.4	/usr/local/pvm3

SCALAPACK-1.7	/usr/local/SCALAPACK

1.2 – CONFIGURATION LOGICIELLE DU CLUSTER : NŒUDS DE CALCUL

Distribution Linux	**Red Hat 7.3**
Kernel	**2.4.18 ou sup**

Nœuds de calcul	noms : node**$i**.domaine.com par défaut : master0.clustal.com

eth0	192.168.1.1(0)**$i**
netmask	255.255.255.0
réseau	192.168.1.0
diffusion	192.168.1.255

partitions du disque dur	
/boot	100 MO
/	3 GO
/tmp	->
swap	RAM x 2

montage nfs	master0:/home sur/home
	master0:/usr/local sur /usr/local

floppy	/dev/fd0

compte root	mot de passe : *****
compte admin	mot de passe : *****

1.3 – SWITCH FAST ETHERNET (pour HP et 3Com manageables)

IP	192.168.1.98
login	login constructeur
mot de passe	mot de passe constructeur
interface web	http://192.168.1.98/

2 – PROCEDURE DE DEMARRAGE ET ARRET DU CLUSTER

*La procédure de **démarrage** du cluster est la suivante :*

1. vérifier l'alimentation électrique
2. brancher le switch Fast Ethernet
3. démarrer le(s) serveur(s)
4. attendre l'écran de login du serveur
5. démarrer les nœuds de calcul
6. travailler

*La procédure **d'arrêt** du cluster est la suivante :*

***option 1** : arrêt par moniteos*

1. passer su dans la fenêtre des commandes parallèles de moniteos
2. sélectionner les nœuds node**n** du cluster qui doivent être arrêtées (ne pas sélectionner le serveur)
3. lancer la commande parallèle shutdown
4. sur les machines SMP : arrêt électrique manuel des machines

***option 2** : arrêt manuel*

1. arrêt des nœuds node**n** (n=0-7) du cluster :
   depuis le serveur :
   su
   rlogin node**n** (n=0-7)
   halt ou shutdown –h now
2. arrêt du serveur
   su
   shutdown –h now
3. sur les machines SMP : arrêt électrique manuel des machines

débrancher le(s) switch(s)

3 – EXPLOITATION DU CLUSTER

3.1 RECOMMANDATIONS GENERALES

Privilégier la connexion utilisateur à la connexion root.

Se connecter sur master0 sur un compte utilisateur

La connexion aux nœuds de calcul doit être réalisée sans authentification (.rhosts en droits 600 dans la racine de l'utilisateur)

Sauvegardes : *a minima*, il est recommandé d'effectuer les sauvegardes des répertoires utilisateurs, c'est à dire de /home sur le serveur

3.2 MANAGEMENT DU CLUSTER

moniteos : monitoring, statistiques et commandes parallèles

se connecter via un navigateur web à l'adresse
http://master0/moniteos/index.html

Se reporter à la documentation spécifique pour l'installation et l'utilisation de moniteos

3.3 DEMARRAGE D'UN NŒUD DE CALCUL APRES LE DEMARRAGE DU SERVEUR

Pour remonter les systèmes de fichiers exportés nfs par le serveur, passer su sur le nœud de calcul et lancer : *bash# mount -a*

3.4 DEMARRAGE D'UN NŒUD DE CALCUL SANS SERVEUR

Si le serveur n'est pas démarré, le montage nfs de /home n'est pas réalisé et les comptes utilisateurs NIS ne sont pas opérationnels. S'il s'avère nécessaire de travailler sur un nœud en station isolée avec des comptes utilisateurs sans pour autant perdre la configuration cluster, procéder de la manière suivante :

- brancher un écran et un clavier (minimum requis)

- se connecter en tant que root ou admin puis su

  1) arrêter le démon ypbind
  2) éditer /etc/fstab

- commenter les lignes (insérer # en début de ligne)
  master0:/home
  master0:/usr/local

Il est alors possible de travailler avec une station pleinement opérationnelle.

nb : le compte admin est créé sur chaque nœud.

4 – EXPLOITATION LOGICIELLE

4.1 MPICH

version 1.2.4

compilation dans /usr/local/mpich-1.2.4

installation dans /usr/local/mpich si un seul compilateur

installation dans /usr/local/mpich-gcc et mpich-autre-compilateur si nécessaire

paramétrage des machines du cluster /usr/local/mpich-1.2.4/util/machines/machines.LINUX et /usr/local/mpich-compilateur/share/machines.LINUX

renseigner avec les nœuds de calcul (important : hostname) ; sur machine SMP : compilation avec l'option --with-comm=shared

4.2 ATLAS

version 3.4.1

installation dans /usr/local/ATLAS

librairies BLAS optimisées dans /usr/local/ATLAS/lib/Linux_P4SSE2_2 pour la configuration nœud.

4.3 LAPACK

version 3.0

installation dans /usr/local/LAPACK

librairie LAPACK optimisée dans /usr/local/LAPACK

4.4 BLACS

BLACS-MPI version 1.0

BLACS-PVM version 1.0

installation dans /usr/local/BLACS

librairies BLACS MPI dans /usr/local/BLACS/lib/

librairie BLACS PVM dans /usr/local/BLACS/lib/

4.5 SCALAPACK

version 1.7

installation dans /usr/local/SCALAPACK

librairies SCALAPACK (MPI, PVM) dans /usr/local/SCALAPACK

4.6 OPEN PBS

version 2.3.16
installation dans /usr/local/OpenPBS_2_3_16
paramétrage des machines du cluster dans /usr/spool/PBS/server_priv/nodes

4.7 LAM MPI

version 6.5.7
compilation dans /usr/local/lam-6.5.7
installation dans /usr/local/lam-mpi
paramétrage des machines du cluster dans
/usr/local/lam-6.5.7/etc/lam-bhost.def et /usr/local/lam-mpi/etc/lam-bhost.def

4.8 PVM

version 3.4.4
installation dans /usr/local/pvm3
paramétrage des machines du cluster dans /home/admin/pvmhostfile

5 – MONITEOS v2.6

5.1 Introduction

5.2 Accessing Moniteos

5.3 Graphic "Monitoring"

5.4 "bMoniteos" Textual monitoring

5.5 The "Parallel Command" tool

5.6 The "PBalineoS" batch system tool

5.7 "PBalineoS" and mpi-ch

5.8 Interactive Jobs

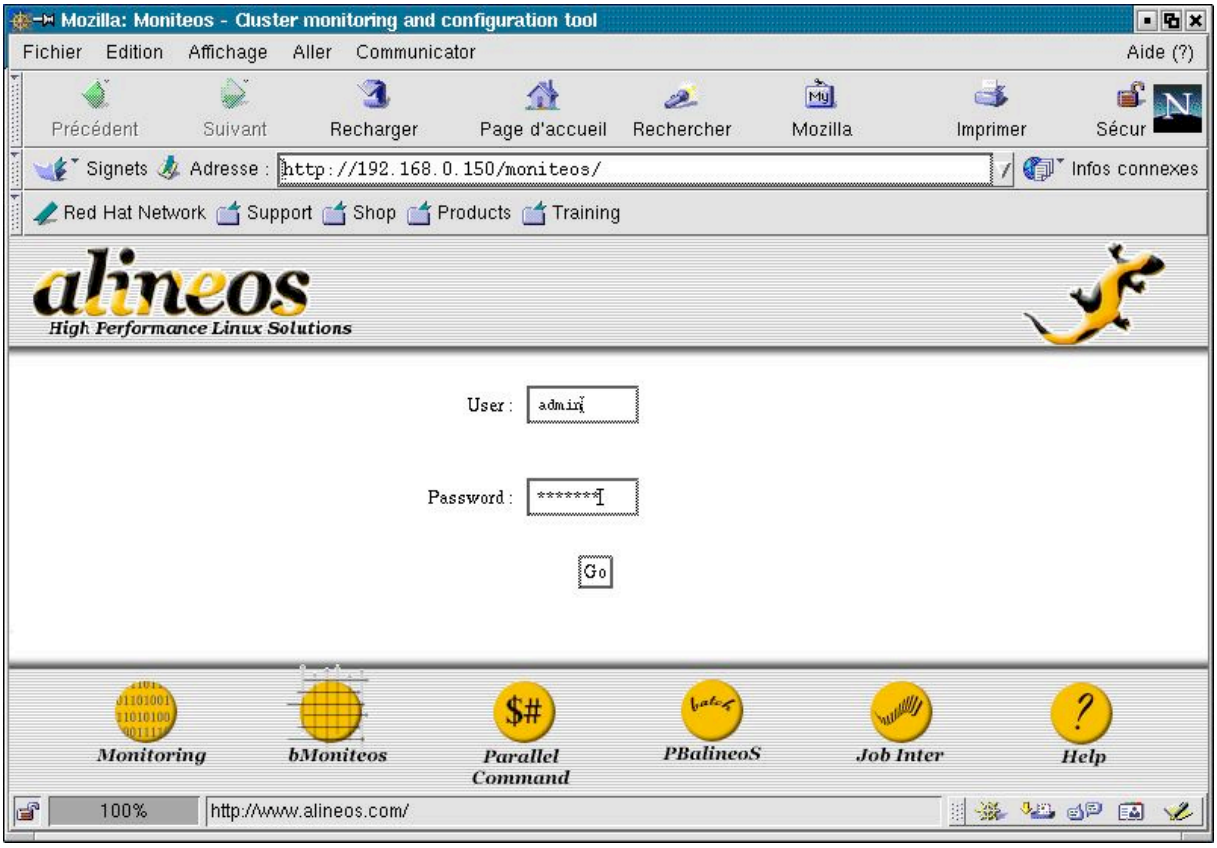
5.9 License agreement

5.1 Introduction

Moniteos is a complete cluster administration tool. It allows you to watch in real-time the cluster load, as a graphic or textual report ; you can also obtain statistics on the utilization of the cluster, node by node. You can, always through the same web interface, execute instructions that will be treated by the nodes you selected. There is no limitation on the number of nodes.

Next - Return to index

5.2 Accessing Moniteos

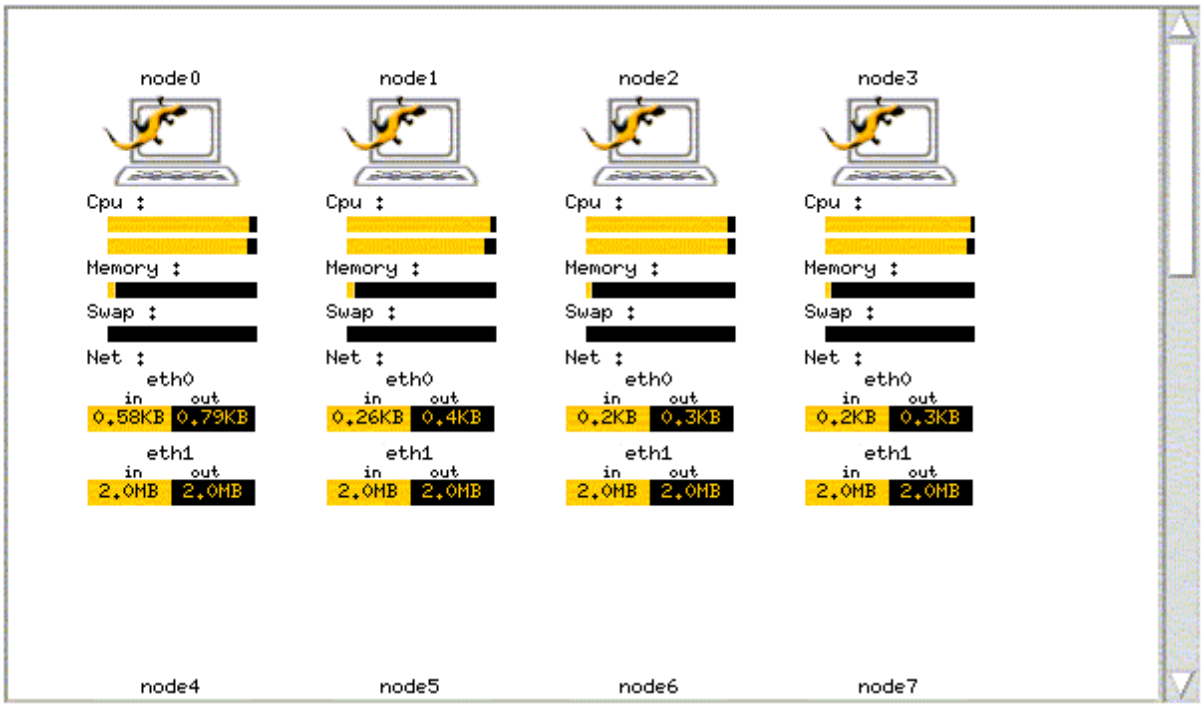


The Moniteos interface is accessed by a web browser : you have to connect to the master node of the cluster, and then, assuming its name is master0, browse to http://master0/moniteos/. This will lead you to Moniteos main page.

Note : Only the help icon is available without log in.

Previous - Next - Return to index

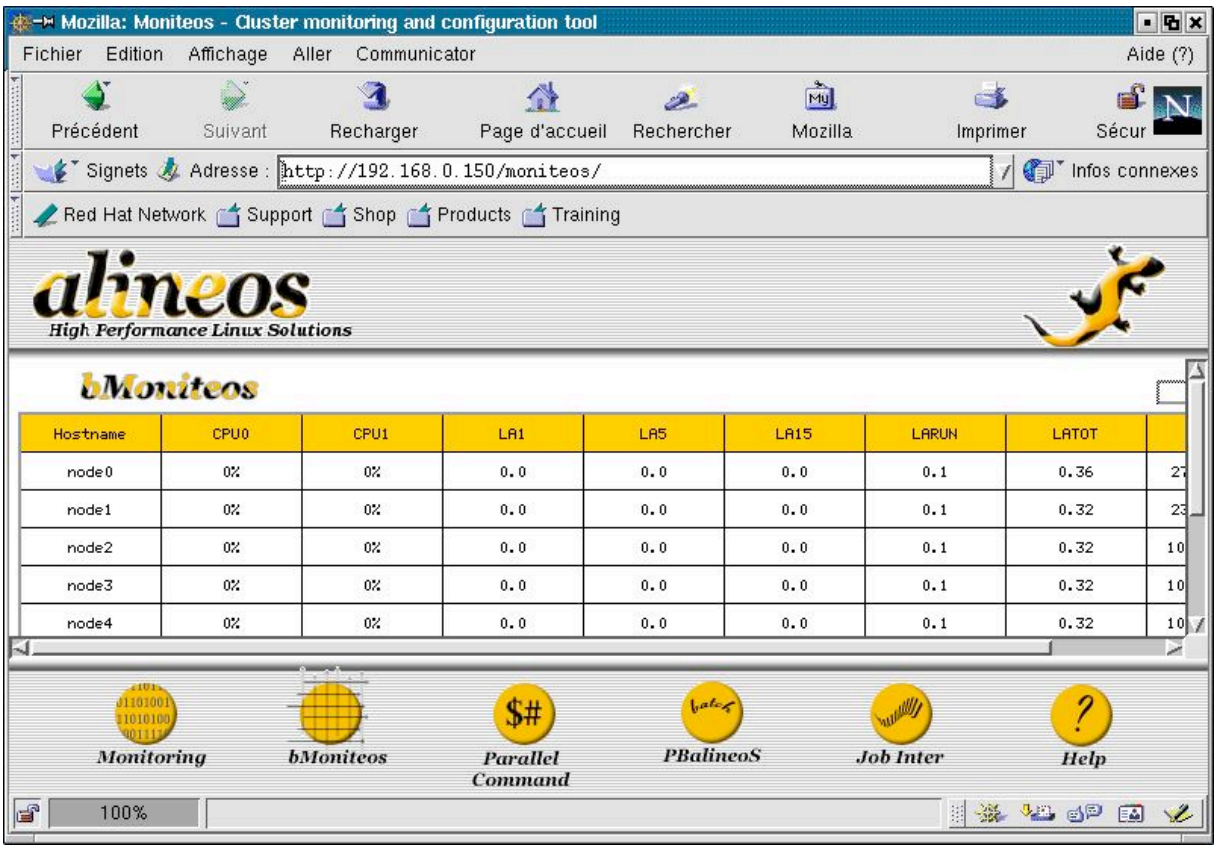
5.3 Graphic Monitoring



This first application is launched automatically by Moniteos at startup. You can access it later by clicking on the first icon, at the bottom of the page. This monitoring applet gives you a synthetic view of the cluster, node by node. Each node has four indicators :
CPU, Memory, Swap, and Network state. If you select one or more nodes by clicking on the screen icon, and
then right-clicking on one of them, you can access statistics of their utilization, from one hour to one month. Note
that if N/A is displayed, the considered nodes could have problems ; verify these nodes' state.

Previous - Next - Return to index

5.4 Textual monitoring

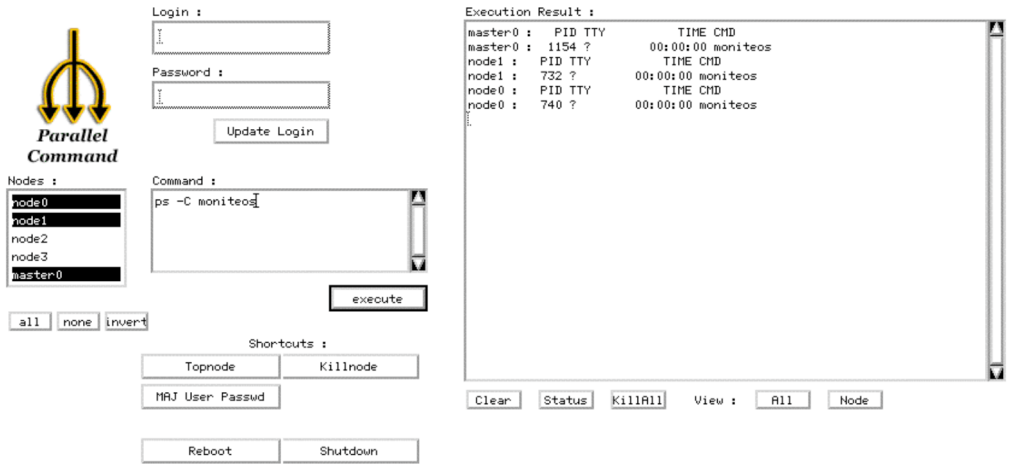

Mozilla: Moniteos - Cluster monitoring and configuration tool
Fichier Edition Affichage Aller Communicator Aide (?)
Précédent Suivant Recharger Page d'accueil Rechercher Mozilla Imprimer Sécur
Signets Adresse : http://192.168.0.150/moniteos/ Infos connexes
Red Hat Network Support Shop Products Training
alineos
High Performance Linux Solutions
bMoniteos
Hostname CPU0 CPU1 LA1 LA5 LA15 LARUN LATOT
node0 0% 0% 0.0 0.0 0.0 0.1 0.36
node1 0% 0% 0.0 0.0 0.0 0.1 0.32
node2 0% 0% 0.0 0.0 0.0 0.1 0.32
node3 0% 0% 0.0 0.0 0.0 0.1 0.32
node4 0% 0% 0.0 0.0 0.0 0.1 0.32
Monitoring
bMoniteos
Parallel Command
PBalineoS
Job Inter
Help
100%

The textual monitoring application is accessed by clicking on the icon "bMoniteos". This application allows you to control the cluster state as a table : each column describes a particular state of the cluster.
We precise here what mean the columns' labels.

CPUx: the mean rate of utilization of the cpu (as a %)
LA1: mean number of processes in running state per minute
LA5: mean number of processes in running state for five minutes
LA10: mean number of processes in running state for ten minutes
LATOT: total number of processes in running state on the system
MEMUSED: amount of used memory on the system
MEMFREE: amount of free memory on the system
MEMSHAR: amount of shared memory on the system (included in MEMUSED)
MEMBUF: amount of buffered memory on the system (included in MEMUSED)
MEMCACH: amount of cache memory used on the system (included in MEMUSED)
SWAPUSED: amount of swap used
SWAPSIZE: amount of swap space on the system
IFxNAME: name of the x network interface
IFxSEND: number of bytes sent by this interface per second
IFxRECV: number of bytes received by this interface per second
IFxCOLL: number of bytes per second which collided

Previous - Next - Return to index

5.5 The "Parallel Command" tool



The Parallel Command application is accessed by clicking on the third icon.

This tool allows you to select a set of nodes (from one node to all) , and to type in a command that will be executed by this set of nodes. In particular, if you need to execute this command under a particular login, fill in the two fields "**Login**" and "**Password**", and then click on "**Update Login**" just below. You can then type your command in the "**Command**" field, and click on "**Execute**" to begin execution. You can see the output of your command either for the whole set of node or node by node. (Note that any command you type must not prompt user for data input while executing.) The "**Status**" button displays in the main window the state of execution of the command, node by node. The "**KillAll**" button stops commands on each node.

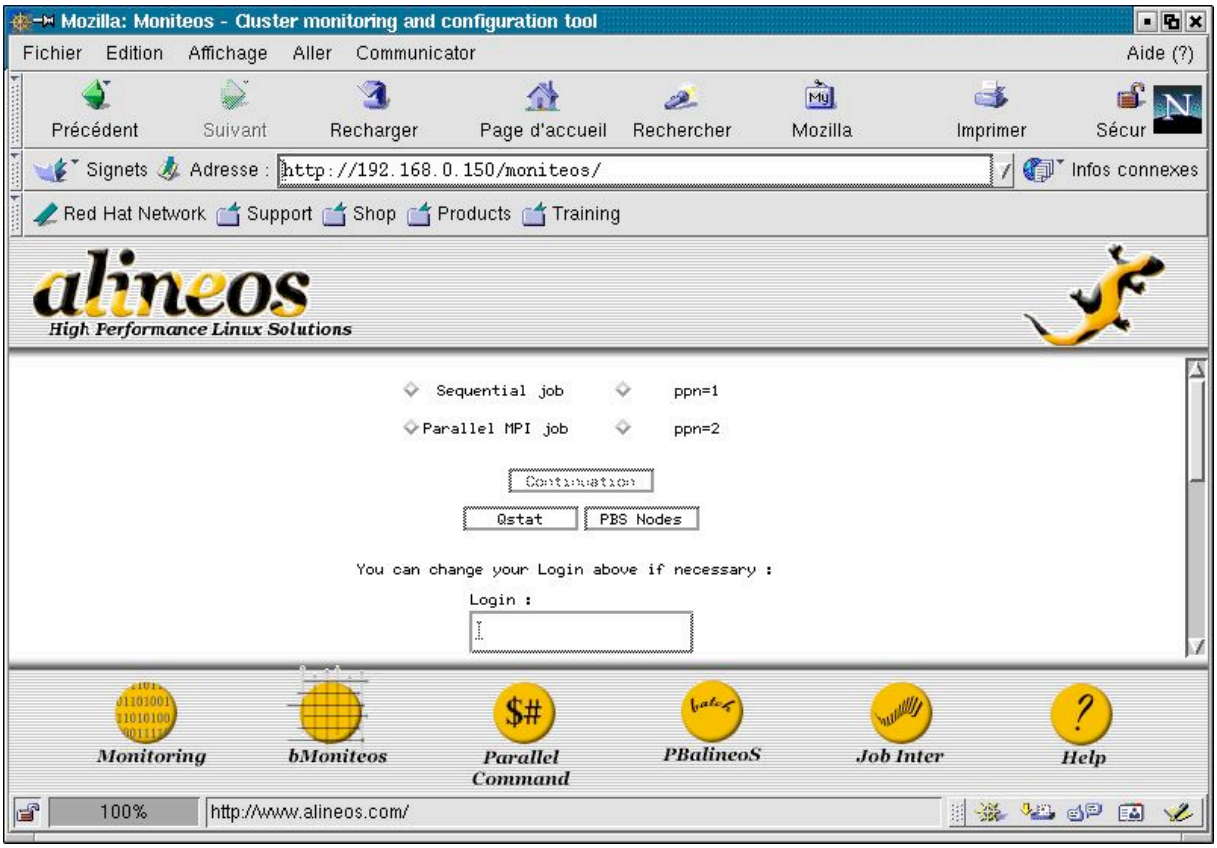
In addition, 5 other buttons were added corresponding to perform programmed actions :

- the **Topnode** button displays - for the selected nodes - the 2 jobs having the most cpu weight
- the **Killnode** button allows the super user to kill a named job on the selected nodes
- the **MAJ User Passwd** allows the super user to propagate the user authentication from the server to the selected nodes. Please remember to create an appropriate .rhosts thereafter
- **Make Boot disk** : it's only to make a boot disk
- **Install Node** : please refer to the PRO version documentation

---

Previous - Next - Return to index

5.5 The "PBalineoS" batch system tool



The PBalineoS application is accessed by clicking on the fourth icon.

This tool is a web interface to OpenPBS. It allows you to submit jobs to the PBS server according to the defined PBS attributes of corresponding nodes and queues. First, you will have to choose what type of jobs you want to submit to :

1) sequential or parallel
2) on ip based network or gm (Myrinet) based network
3) with 1 processor per node or 2 processors

Then a new window opens.

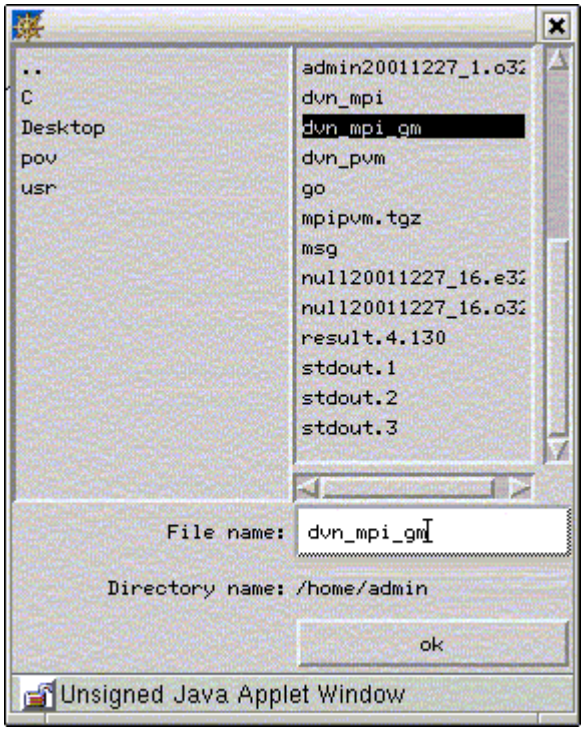
Previous - Next - Return to index

To submit a job, you need to **give the following informations** :

- a job name (optional)
- the name of the selected queue
- the number of nodes you need
- the maximum time of the job (00:00:00 means no limit)
- do you want to merge the stdout with the stderror ?Yes or No...
- an email address to send messages when jobs starts, ends or aborts
- the filenames of stagein and stageout files

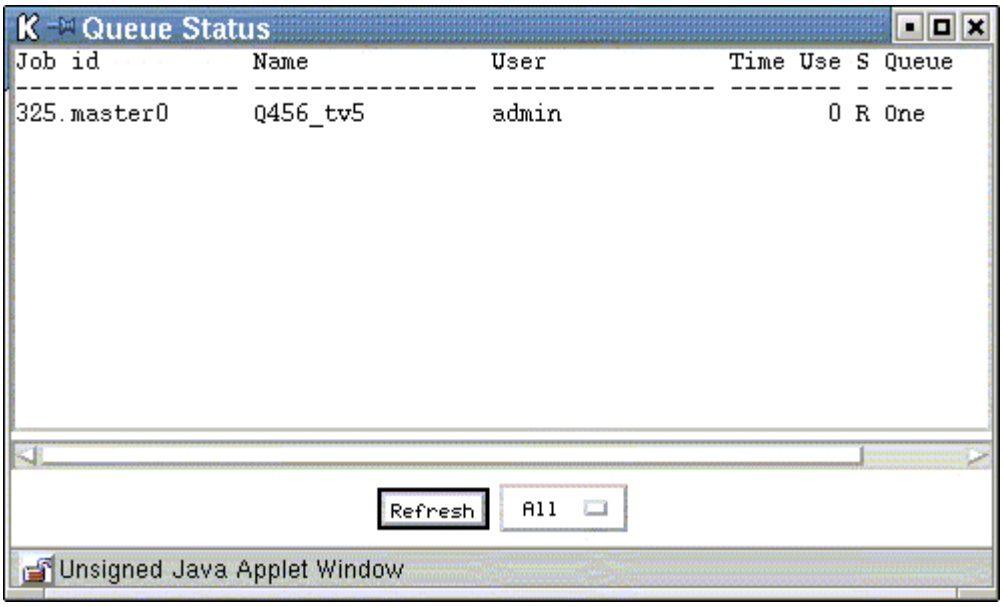
and then, select a script file by clicking the **Load Script button**.

Previous - Next - Return to index



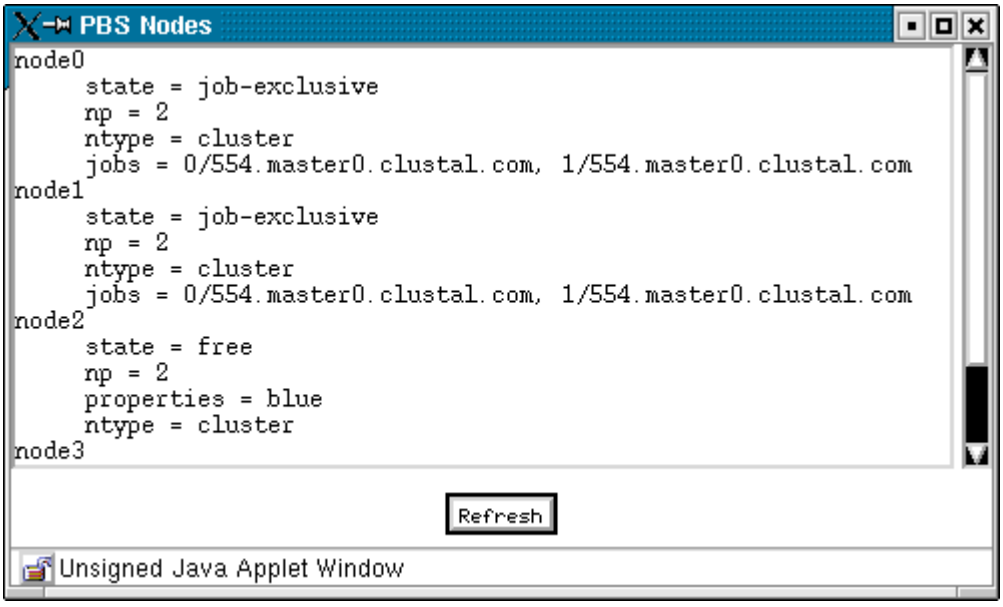
You will browse in your directories tree and select a file containing the appropriate instructions for the job. The script is then loaded in the corresponding applet's window. In this window, you always can modify the pre-loaded instructions. After this step, you will create the PBS script by clicking the **Create Script button** and all the instructions then appear in the second applet's window. Once again, you can modify in this window all the instructions, including PBS directives. Finally, submit your job by clicking the **Submit button**.

Previous - Next - Return to index



When running PBS, you at every time can select a queue and then click on the **Qstat button** to display the status of the selected queue.

You can also click on the **PBS Nodes button** to display the status of all the node.



Note : **Qstat button** and **PBS Nodes button** are also available from the access page of Pbalineos.

Previous - Next - Return to index

5.5 Parallel jobs, PBalineoS and MPI

To use the full features of mpi-ch implementation to OpenPBS via PBalineoS, you only need to give following arguements to your command line :

for mpi-ch over IP :

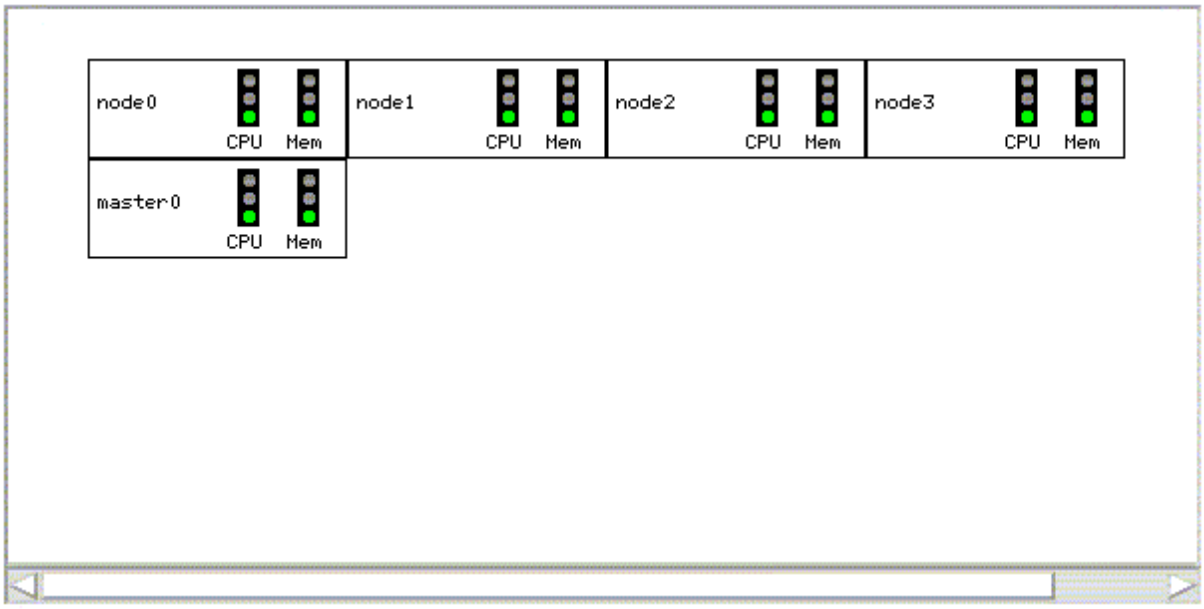
-machinefile $confile -np $nnodes

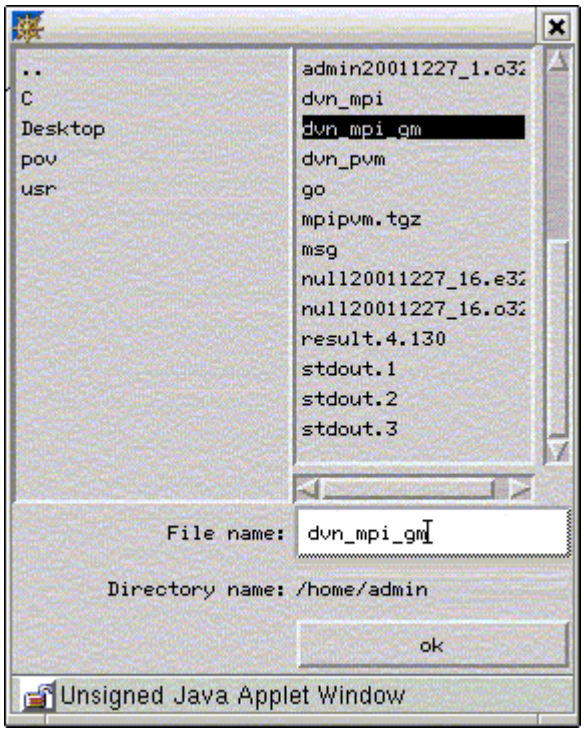
and for mpi-ch over GM :

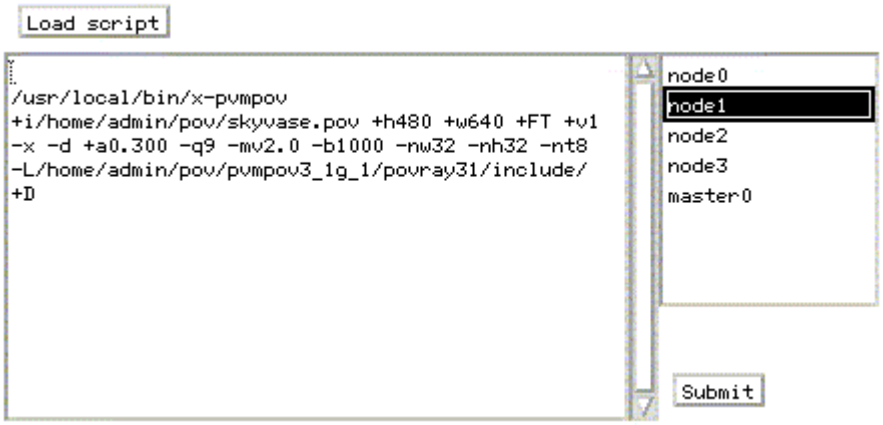
--gm-f $confile –np $nnodes

Previous - Return to index

5.8 Interactive Jobs







The Interactive Job Submission application is accessed by clicking on the fifth icon.

This tool is a web interface which allows you to submit jobs to a selected node with respect to its cpu and memory load.

To submit a job, you only need to do following :

- a job name (optional)
- select a script file by clicking the **Load Script button**. You will browse in your directories tree and select a file containing the appropriate instructions for the job. The script is then loaded in the corresponding applet's window. In this window, you always can modify the pre-loaded instructions.
- After this step, you will choose the node you want to submit the job to.
- Finally, submit your job by clicking the **Submit button**.

Previous - Return to index

5.9 Moniteos License Agreement

NOTICE : PLEASE READ CAREFULLY THIS DOCUMENT BEFORE USING THE SOFTWARE.
BY USING THE SOFTWARE, YOU ARE AGREEING TO BE BOUND BY THE TERMS OF THIS LICENCE.
IF YOU DO NOT AGREE TO THE TERMS OF THIS LICENCE, PLEASE DO NOT USE THE SOFTWARE.

You, as a Licensee under this Agreement, are hereby granted a limited, irrevocable, nontransferable and nonexclusive license to use the Software subject to the restrictions and other terms within. That use must be only by the Licensee. Any (complete or partial) reproduction or distribution of the software or its source code is strictly prohibited. Licensee may not rent, lease, loan, electronically transfer the Software and its source code to others. The Licensee may modify the source code for his own use.

YOU ACKNOWLEDGE THAT YOU HAVE READ THIS AGREEMENT AND AGREE TO BE BOUND BY ITS TERMS. YOU FURTHER AGREE THAT IT IS THE COMPLETE AND EXCLUSIVE STATEMENT OF AGREEMENT BETWEEN US WHICH SUPERSEDES ANY PRIOR AGREEMENT, ORAL OR WRITTEN, ANY PROPOSAL AND ANY OTHER COMMUNICATIONS BETWEEN US RELATING TO THE SUBJECT MATTER OF THIS AGREEMENT.

ALINEOS
14 bis rue du Marechal Foch
77780 Bourron-Marlotte
France
http://www.alineos.com

6 – CHECK LIST D'INSTALLATION CLUSTER

- **Vérification des partitions :**
  - Sur le frontal :
    df –k
  - Sur les nœuds :
    ./execnode " df –k " -t 7
    Vérifier le montage de /usr/local et /home
  - Vérifier le swap :
    free | grep Swap

- **Vérification de versions noyau :**
  - Sur le frontal :
    uname –a
  - Sur les nœuds :
    ./execnode " uname –a " -t 7

- **Vérification des taux d'accès disques :**
  - Sur le frontal :
    hdparm –tT /dev/hda
  - Sur les nœuds :
    ./execnode " hdparm –tT /dev/hda " -t 7

- **Vérification de la configuration IP :**
  - Vérifier la configuration IP cluster.
  - Paramétrer l'IP externe, DNS, Gateway et nom de machine (/etc/hosts).

- **Vérification des taux d'accès réseau MPI**
  - Script de lancement de l'exécutable d'exemple de mpich systest dans /home/admin : mpisystest.sh

- **Moniteos**
  - Verification version (Logo High Performance Linux Solutions)
  - Monitoring :
    Vérifier que tous les noeuds sont bien en ligne.
  - BMoniteos :
    Vérifier que les données sont bien renseignées pour tous les noeuds.
  - Parallel Command :
    Lancer une commande sur plusieurs noeuds : exemple echo $$.
  - Pbalineos :
    qstat, PBS nodes, lancement du calcul de pi depuis le script /home/admin/mpicpi-pbalineos.sh.

- **OPEN PBS**
  - Lancer trois fenêtres xterm.
  - Dans le premier taper qsub –I –l nodes=2 :ppn=2
  - Dans le second taper qsub –I –l nodes=<X> :ppn=2 avec X de sorte que X+2=Nombre total de nœud du cluster+1
  - Dans le troisième taper qstat –n et vérifier l'état des jobs (le second doit être mis en attente).

- **ATLAS**
  - Librairies disponibles dans /usr/local/ATLAS/lib/<ARCH>
  - Installation de gcc 2.95.3 permettant une meilleure optimisation. Utilisable par : source /home/admin/.gcc-2.95.3

- **LAPACK**
  - Librairie dans /usr/local/LAPACK

- **BLACS**
  - Librairies MPI et PVM dans /usr/local/BLACS/LIB

- **SCALAPACK**
  - Librairies MPI et PVM dans /usr/local/SCALAPACK

- **PVM**
  - Configuration de nouveaux comptes : 4 lignes à rajouter au fichier .bashrc (Cf. /home/admin/.bashrc) :
  - Test depuis le node0 dans /home/admin :

    pvm pvmhostfile

    puis au promt :

    pvm> conf -- renvoie la configuration du cluster

    pvm> spawn -> hello

    pvm> halt

- **LAM**
  - Configuration de nouveaux comptes : 3 lignes à rajouter au fichier .bashrc (Cf. /home/admin/.bashrc)
  - Test depuis le node0 dans /home/admin

    recon -- test la configuration pour LAM

    lamboot –s lamhostfile -- lance LAM avec la configuration du fichier

    lamnodes -- affiche la configuration

    /home/admin/TestLam.sh -- script qui lance deux programmes d'exemples -- LAM : ring et cpi (calcul de pi)

    lamhalt

    Deux fichiers résultats se trouvent dans /home/admin/ :
    - lampitest.txt
    - lamringtest.txt